名古屋外国語大学外国語学部紀要第50号(2016.2)

ツイートからの観光ルート抽出

若　山　公　威

1.　はじめに

Twitterはツイートという140文字以内のテキストを投稿するマイクロブログであり、ユーザが自身の体験したことや身の回りで起きた出来事などを投稿している。これらの情報から特定地域に関するものを取り出すことで、その地域での様々な出来事が分かる。また、旅行者がどこで何をしたかという情報も多く含まれており、観光地の様子がリアルタイムで投稿されるため、他の旅行者に有益な情報となる。しかし、このような情報を整理し有効活用する技術の研究は十分には進んでいない[1]。

ツイートから観光地の様子を取り出す場合、その記述がどの場所に関するものであるか把握する必要がある。地名を含んでいない場合や、地名を含んでいても必ずしもその場所にいる訳ではないためである。ジオタグが付いているツイートも存在するが、全ツイートのうち約0.18%のみしかない[2]。したがって、多くの情報を取得するには、ジオタグなしツイートも対象とする必要があるため、ツイート本文のみから発信者の位置推定を行う研究が多く行われている[3]。ユーザの位置情報が分かることで、そのユーザが発信するツイートからその場所に関する最新の状態を知ることができるだけでなく、そのユーザに対して、ユーザの趣味嗜好や属性に応じた周辺の店や観光名所などの情報を提供することも可能となる。

観光地では複数の場所を順に訪れることが多く、そのルートを特定する

ことも重要である。観光ルートを抽出できると、場所を明記されていないツイートに関しても、前後関係からどの場所に関して投稿されたツイートであるかを把握することができるようになり、観光情報を整理することに役立つと考えられるためである。また、他の旅行者に観光地を推薦する際に、訪れる順番も重要な要素となる。

本研究では、ジオタグなしツイートにおいて、観光ルートを抽出する方法を提案し、実際のツイートを用いて有効性を検証する。

2. 関連研究

小原ら[4]は、ジオタグなしツイート本文から地域連想語や観光に関するツイートに用いられやすい単語を用いて観光情報を抽出する方法を提案している。具体的には「なう」、「到着した」、「楽しかった」などの単語を含むかどうかのパターンマッチングを行っている。観光ルートの抽出は行っていない。

中嶋ら[5]は、観光地付近でつぶやかれたジオタグ付きツイートからユーザが旅行者かどうか判別し、その旅行者がつぶやいた観光に関するツイートを収集し、観光ルートを推薦する方法を提案している。観光地周辺から投稿されたジオタグ付きツイートのうちFoursquareとInstagramを利用したツイートや、ジオタグが付いていないツイートのうち「清水寺なう」のように観光地名とともに「なう」が書かれているものを旅行者のツイートとして扱っている。新井ら[6]は、ジオタグが付与されていないツイートに関して、観光地を訪れたとみなされる表現を増やしている。この方法では、過去の観光を振り返るツイートも対象としてしまう場合がある。

鬼塚ら[7]は、観光情報を対象として現地性判断を行っている。現地性とは、投稿者がその場にいて体験したことに関するものかどうかを識別することである。その場所から発信されたものの場所と関係ない雑談であったり、別の場所からその場所について述べていたりするツイートには現地性

がない。現地性が判断できれば、ツイートから観光情報として適切なもののみが収集可能となる。鬼塚らの提案方式では、まず時間帯や特定の言葉を含むかどうかといったルールによりツイートをフィルタリングし、このルールで判定できなかったものについては機械学習により判定している。そして、「清水寺」を含むツイートを対象に実験を行っている。

現地性に関する研究は、観光以外に関しても行われている。蛭田ら[8]は、位置情報付き発言の中でも、現在の場所で起きた出来事・状況などに誘因されて発言されたものを「場所誘因型位置情報付き発言」と定義し、特定のキーワードが含まれているかどうか調べることで判定を行っている。この研究では、ジオタグが付いていないツイートは対象としていない。宮部ら[9]は、場所依存記録と呼ばれる場所と密接に関連したツイートであるかどうかをSVMにより判定している。入力は1-gram、2-gramの形態素のみを用いている。鬼塚らの方式を併用することにより更なる向上が期待できる。

本研究では各ツイートの現地性は扱わずに、現地に着いたことを述べているツイート群からルート抽出のみを扱う。ルート抽出後、各ツイートから現地性があるものとないものとを分別して観光情報を取り出していく予定である。

3. 提案手法

ここでは、ユーザ本人が移動した経路を含むツイートを収集し、訪れた地名を取り出す方法について説明する。ツイートに地名が含まれていても、「『観光地名』へ行きたいな」といった願望であったり、過去の旅行について回想していたりする場合がある。本研究では、ツイート本文の表現からパターンマッチングにより、現地にいる可能性が高いものを取り出す。

まず、Twitter API[1]を利用して、任意の観光地名を含むツイートを収集す

[1] https://dev.twitter.com/

る。次に、各ツイートについて、投稿したユーザのタイムラインを取得する。これらのうち、現地を訪れたとみなせる以下の表現が含まれるツイートから地名を取り出す。

(1) 地名+(「に」)+「着いた/ついた/着きました/つきました/到着」
(2) 地名+(「に/まで」)+「来た/きた/来てみた/きてみた/来てみました/きてみました」
(3) 地名+「なう/なぅ/なーう/な〜う/なう〜/わず」
(4) 地名+「!/ー」
(5) 本文に地名が1つ含まれ、写真が添付されている

項目(3)の「なう」は、ツイートで場所や動作とともに頻繁に用いられ、現在の場所や自分が行っていることを表すものである。「わず」は、その場所を訪れた直後に利用されることが多いため、今回用いることとする。

観光地では写真を撮り投稿することが多いため、項目(5)を入れている。ただし、複数の観光地を訪れた後にまとめて複数の写真を投稿することもあるため、本文に地名を1つのみ含むものを対象としている。

同じ日であっても、一日の移動を振り返るツイートは訪問した詳細な時間が分からないため対象外とする。このような振り返りツイートでは1ツイート内に訪問先が複数書かれていたり、短時間に投稿される複数ツイートに一日分の訪問先を書いたりすることが多い。本研究ではこの特徴を利用して、直前に取り出された地名との直線距離と投稿時間差から移動速度を求め、閾値以上の場合は破棄することで、振り返りツイートを削除する。距離を求める際に同名地名が複数ある場合は、検索地名から最も距離が近い地名を選択する。

以上のルールで取り出した観光ツイートから地名のみをツイート投稿日時とともに順に並べていくことで、観光ルートを抽出する。地名としては、市町村名、寺社名、遊園地などが含まれる。また、サービスエリアや駅など観光地以外の場所も含まれる。なお、同じ地名が連続して抽出された場合は、最初のもののみを選択している。

4. 実験

4.1 使用データ

提案方法を評価するために実験を行った。ここでは、実験に使用したデータについて説明を行う。まず、PHPプログラムを用いてTwitter APIから、2015年8月1日のツイートのうち本文に「清水寺」を含むものを取得した。Botによる自動投稿の影響を小さくするため、Twitterクライアント名に“Twitter for”を含むもののみとした。該当するツイート数は280であった。次に、各ツイートについて、投稿したユーザのタイムラインを該当ツイート前後24時間分ずつ取得した。リツイートは除いた。また、宣伝ツイートを除くため、鬼塚らの研究[7]を参考に本文に100文字以上含むツイートも削除した。最終的に残ったツイート数は13068である。これらのツイートについて、ツイートID、ユーザID、投稿日時、本文、添付画像を取得して利用した。

地名辞書として、次のデータから地名と位置情報（緯度・経度）を取り出したものを用意した。

- 「駅データ.jp」[2]の駅名データ　9049件
- GeoNLP[3]の都道府県および市町村名データ　4562件
- 2015年7月3日時点のWikipediaのうち位置情報が含まれているキーワード　80029件

ツイート本文から、ハッシュタグ、URL、RT、@などをノイズとして除去したのち、MeCab[4]を用いて形態素解析を行った。MeCab辞書には前述のWikipediaキーワードを追加した。形態素解析の結果、地名となった単語に関して、3節のルール（1）から（5）に該当するか順にチェックしていっ

[2] http://www.ekidata.jp/

[3] https://geonlp.ex.nii.ac.jp/

[4] http://taku910.github.io/mecab/

た。該当した場合、地名と投稿時間をルートに追加していった。また、地名辞書から位置情報を取り出し直前の場所からの直線距離を求め、投稿時間の差を用いて移動速度を求めた。ただし、海外の地名は対象外とした。移動速度の閾値は自動車での移動を想定して時速60kmとした。

4.2 実験結果

3節の(1)から(5)、各ルールで取り出されたツイート数を表1に示す。複数のルールに一致している場合は、上位のルールを示している。75%ほどが写真により抽出されていることが分かる。

表1　ルールごとの該当ツイート数

ルール番号	該当したツイート数
(1)	25
(2)	24
(3)	32
(4)	36
(5)	351

取り出されたルートのうち、1つ以上の地名が含まれているものは123であった。このうち、5か所以上地名が含まれている15ルートについて、次の式を用いて地名単位での適合率、再現率、F値を求めた。

$$\text{適合率} = \frac{TP}{TP+FP} \qquad \text{再現率} = \frac{TP}{TP+FN} \qquad \text{F値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

TPは正しく抽出できた地名数、FPはそこにいないにも関わらず抽出してしまった地名数、FNは現地にいたにも関わらず抽出できなかった地名数である。正解データはツイート本文を人手で確認して作成した。結果を表2に示す。適合率は誤抽出の少なさ、再現率は抽出漏れの少なさを示すため、今回は誤抽出より抽出漏れの方が多いことが分かる。

正しくルートを抽出できたあるユーザのツイート群を表3に示す。下線

表2　評価結果

適合率	再現率	F値
88.0%	73.6%	80.2%

表3　地名抽出成功例

番号	投稿日時	ツイート本文
1	2015-08-01 09:23:32	滋賀県の長浜城に到着しました。 http://t.co/d07juq2X0Z
2	2015-08-01 11:47:41	10円玉で有名な平等院鳳凰堂に到着です。 http://t.co/Hd3vXThFGS
3	2015-08-01 12:44:47	こんなところに任天堂。 http://t.co/90MIpqJ63i
4	2015-08-01 14:51:29	東大寺、凄い大きい。 http://t.co/DF8CclpfUV
5	2015-08-01 16:42:59	金閣寺、凄い綺麗。 http://t.co/EIhlYpAB8B
6	2015-08-01 17:59:30	本日最後の観光地、清水寺。 友人に「清水寺」があだ名だった人がいるので、前から行ってみたかった場所です。 http://t.co/RETXBvcyQg
7	2015-08-02 07:40:56	たこ焼きのレベルが高い大阪は、たこ足配線のレベルも高い。 http://t.co/ZE03f99JEN

表4　移動速度による地名フィルタリング成功例

番号	投稿日時	ツイート本文
1	2015-08-01 08:19:00	太秦の映画村、金閣寺、龍安寺、北野天満宮 http://t.co/jUgEoUP22y
2	2015-08-01 08:21:59	清水寺、二年坂 http://t.co/6h4hLO01Ew

が引かれている個所が抽出された地名である。URLが書かれているツイートは画像が添付されているものである。直前の場所からの移動速度を用いることにより、振り返りを行うツイートの削除が行えている例を表4に示す。2つのツイートとも前日の観光地での写真を投稿しているもので、短時間のうちに距離が離れた地名が表れているため、ルート地名抽出の対象とはなっていない。

誤検出した地名について原因をまとめた結果を表5に示す。

まず、原因が辞書に関するものについて述べる。最も多い原因である「地名切り出し失敗」は、地名がMeCab辞書に登録されていないことが原因で、形態素解析による切り出しが失敗したものである。今回は、Wikipediaのキーワードは登録したものの、駅名、都道府県名、市区町村名のデータは、MeCabの標準辞書でカバーできると考えたため登録しなかった。これらを登録するとともに、特に表6の1番のような観光地を充実させる必要がある。「地名の別名」は、例えば表6の3番のように「西本願寺」を「西本」と記述していたり、2番のようにローマ字で書かれていたりするものであり、これもMeCab辞書の充実が必要となる。ツイートでは省略や俗語の利用が多いため、通常の辞書では対応が難しい。この対策として、FoursquareなどのSNSで利用されている地名データを利用することにより、観光地や一般に使われている別名などを増やすことが有効と考えられる。「地名の位置情報なし」は、地名がMeCab辞書には含まれていても、地名辞書に含まれていなかったため位置情報を取り出すことができなかったものである。MeCab辞書に加えて、地名辞書の充実も必要である。これらの対策により抽出漏れを防ぐことで、表2の再現率の向上につながると期待できる。

「写真添付ツイート」は、写真が添付されているツイート本文に現地とは別の地名が書かれていたもの、あるいは、本文には該当地名が書かれていても現地性がないものである。今回は写真付きツイートの場合は地名が書いてあれば現地と判断しているが、ツイート本文の解析が必要と考えられる。表1から、写真が添付されていることで現地性を決定しているものが多いため、この対策により精度の向上が期待できる。

「速度超過」は、振り返りツイートを削除するための移動速度閾値を超えたものである。おもに、新幹線や高速道路利用による、京都市外から市内への移動の際に発生している。直近のツイートから移動手段が判断できる場合は、閾値を変動させることで改善が可能と思われる。

「パターンマッチング非該当」は、表6の5番のように現地に着いた表現であるにも関わらず、今回の抽出ルールに含まれていなかったものである。

現地性があるツイート本文を調査することにより、さらなるルールの充実が必要である。

「本文100文字超過」は広告ツイートを削除するルールにより、正解ツイートが削除されてしまったものである。例として、表6の4番があげられる。広告ツイートの本文を詳しく調査して、文字数ではなく内容により削除するために、新たなルールを決める必要がある。

表5　失敗の種類ごとの該当地名数

失敗の種類	地名数
地名切り出し失敗	11
写真添付ツイート	10
地名の別名	6
速度超過	4
パターンマッチング非該当	4
地名の位置情報なし	4
本文100文字超過	2

表6　地名抽出失敗例（すべて別ユーザ）

番号	投稿日時	ツイート本文
1	2015-08-01 10:54:08	哲学の道で哲学 http://t.co/v3MLvNi1vx
2	2015-08-01 16:10:25	Ryoan-ji http://t.co/SQ4cEDpeat
3	2015-08-01 16:33:34	劇場版　西本の願い http://t.co/nec7oOB4NZ
4	2015-08-01 17:33:37	ラストは清水寺！ 舞台を下から撮りました。…え？ 舞台から撮った景色？ いやあ、撮れませんでしたわ。 ネクストコナンズヒ〜ント！「高所恐怖症」http://t.co/5Rs2gl7RiT
5	2015-08-02 05:07:02	@yomi0602 あ、お好み焼きも食べなきゃね！ 一人旅になった分、自由にいろんなところ行けるから楽しみ（^o^） 愛知県突入なう

5. まとめ

本研究では、ジオタグなしツイートにおいて、本文のパターンマッチングにより観光ルートを抽出する方法を提案し、実際のツイートを用いて実験を行い評価した。その中から一部のデータを調査した結果、F値は80.2%となり有効性を確認できた。また、直前の移動先からの距離と投稿時間から移動速度を算出することにより、過去の振り返りツイートの削除が可能となった。今回は地名が取り出されたツイートのうち75%が、写真が添付されていることにより現地にいることが判別された。地名取り出しの失敗原因としては、辞書不備による地名切り出しミスが最も多かった。次に多い失敗原因は、写真付きツイートでの地名が現地のものでないためであった。これらの対策が必要なことが分かった。

今後は、地名辞書を充実させるとともに、写真が添付されているツイートの本文のパターンを調査し、実験対象ツイート数を増やして提案方式の有効性を確認する。観光ルートが正しく取り出されるようになったのちには、ツイートから観光地で見たことや行ったことを取り出し、観光地の情報としてまとめていく予定である。

参考文献

［1］難波英嗣: 観光情報の自動編纂, 知能と情報, Vol. 26, No. 1, pp.9-15 (2014).
［2］橋本康弘, 岡瑞起: 都市におけるジオタグ付きツイートの統計, 人工知能学会誌, Vol.27, No.4, pp.424-431 (2012).
［3］Zhiyuan Cheng, James Caverlee, Kyumin Lee: You are where you tweet: A content-based approach to geo-locating twitter users, In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), pp.759-768 (2010).
［4］小原基季, 森田和宏, 泓田正雄, 青江順一: Twitter本文を用いた観光情報抽出及び分析システムの構築, 人工知能学会第29回全国大会講演論文集（2015）.
［5］中嶋勇人, 新妻弘崇, 太田学: 位置情報付きツイートを利用した観光ルート

推薦，情報処理学会研究報告，データベース・システム，Vol.2013-DBS-158, No.28, pp.1-6 (2013).
［6］新井晃平，新妻弘崇，太田学: Twitterを利用した観光ルート推薦の一手法，第7回データ工学と情報マネジメントに関するフォーラム論文集（2015）.
［7］鬼塚友里絵，嶋田和孝: 前後文脈を考慮したTweetの現地性判断，信学技報，Vol. 114, No. 81, NLC2014-5, pp. 23-28 (2014).
［8］蛭田慎也，米澤拓郎，徳田英幸: 場所誘因型位置情報付き発言の検出と可視化，情報処理学会論文誌，Vol.54, No.2, pp.710-720 (2013).
［9］宮部真衣，北雄介，久保圭，荒牧英治: マイクロブログから場所依存の様相記録を抽出する：“100ninmap”プロジェクトによる街歩きイベントの実施と応用，言語処理学会第20回年次大会発表論文集，pp.420-423 (2014).